Section 3.0 Statistical Methods for Establishing Baseline Conditions and Setting Discharge Limits at Remining Sites

3.1 Objectives & Evaluation of Statistical Methods

The Rahall amendment, CWA Section 301(p), states in part:

> *(2) LIMITATIONS. - The Administrator or the State may only issue a permit pursuant to paragraph (1) if the applicant demonstrates to the satisfaction of the Administrator or the State, as the case may be, that the coal remining operation will result in the potential for improved water quality from the remining operation but in no event shall such a permit allow the pH level of any discharge, and in no event shall such a permit allow the discharges of iron and manganese, to exceed the levels being discharged from the remined area before the coal remining operation begins.*"

EPA has promulgated the Coal Remining Subcategory (40 CFR Part 434.70) consistent with the requirements and intent of the Rahall amendment. The regulations for the Coal Remining Subcategory establish a standardized statistical procedure for determining baseline pollutant loadings and pollutant loadings during remining for net acidity, solids, iron, and manganese in pre-existing discharges. These statistical procedures are codified in Appendix B to Part 434 and are intended to identify increases (during remining) of discharge pollutant loadings above the baseline levels.

EPA has interpreted "levels" to mean the entire probability distribution of loadings, including the average, the median, and the extremes. It follows that if P percent of loadings did not exceed some number Lp during baseline, then no more than P percent should exceed Lp during and after remining. For example, if during the baseline period, 95 percent of iron loadings are $\leq$ 8.1

lbs/day and 50 percent are ≤ 0.3 lbs/day, then during and after remining the same relationships should hold true.

The objective of Section 3 is to provide statistical procedures for deciding when the pollutant loadings in a discharge exceed the levels of baseline. These procedures are intended to provide a good chance of detecting a substantial, continuing state of exceedance, while reducing the likelihood of a "false alarm." The procedures (or the numbers calculated from them) are also referred to here as "triggers."

In developing these procedures, EPA considered the statistical distribution and characteristics of discharge loadings data from pre-existing discharges, the suitability of parametric and nonparametric statistical procedures for such data, the number of samples required for these procedures to perform adequately and reliably, and the balance between false positive and false negative decision error rates. EPA also considered the cost involved with sample collection as well as delays in permit approval during the establishment of baseline, and considered the potential that increased sampling could discourage remining. In order to sufficiently characterize pollutant levels during baseline determination and during each annual monitoring period, Appendix B to Part 434 requires that the results of a minimum of one sample be obtained per month for a period of 12 months.

The procedures described below will provide limits for both single observations and annual averages. This is intended to provide checks on both the average and extreme values. There is a need to take into account the number of observations used to determine compliance when setting a limit or when otherwise determining compliance with baseline. For example, the collection of a greater number of samples from a discharge will reduce the variability of the average level (provided that samples are distributed randomly or regularly over the sampling year). Accordingly, the statistical procedures described here take into account the amount of data in an appropriate fashion.

Use of a statistical decision procedure should result in suitable error rates. Technically these are usually referred to as the rate alpha (α) for Type I errors and the rate beta (β) for Type II errors. The error of concluding that an exceedance has occurred when the discharge is exactly matching the baseline condition is intended to happen with probability α. Alpha can be characterized as the maximum "false alarm rate." When the discharge level is substantially less than baseline, the probability of making this error is expected to be very low. The error of concluding that no exceedance has occurred, when the discharge has in fact exceeded baseline levels, is intended to happen with probability β. Power (π) equals 1-β. Power can be defined as the probability that a statistical decision procedure will declare that remining loadings exceed baseline loadings when there really has been an increase, or as the rate of giving correct alarms.

When many decisions will be made, the overall error rate is a concern. For example, the single-observation triggers described below will be applied every month during remining; the annual triggers will be applied every year. In evaluating statistical methods, EPA considered the overall or cumulative decision error rates during a five-year period of compliance monitoring.

The degree of serial correlation of the data will influence the decision error rates of statistical procedures. There is significant, positive serial correlation of flow, concentration, and loading in mine discharges over periods of 1 to 4 weeks, that is, sequential samples are correlated with each other (U.S.E.P.A., 2001a, 2001b). Also, estimates of the variance, used in parametric statistical procedures, are inaccurate unless adjusted for autocorrelation. (Loftis and Ward, 1980; U.S.E.P.A., 1993). Such adjustments require an estimate of the autocorrelation coefficient. However, one cannot reliably estimate site-specific autocorrelation from small samples (e.g., n=12). Using long-term datasets for pre-existing discharges at abandoned mines and at remining sites, EPA estimated the first-order serial correlations (at a monthly interval) for flow and for iron, manganese, and acidity loadings. The estimates fell mostly in the range 0.35 to 0.65, with a central tendency just below 0.50 (U.S.E.P.A., 2001b).

EPA evaluated parametric and nonparametric statistical procedures for characterizing the baseline level and determining compliance with the baseline level (U.S.E.P.A., 2001c). For the

evaluation, EPA simulated discharge loadings data. These data had realistic statistical properties, resembling actual discharge loadings in terms of distribution and serial correlation (U.S.E.P.A., 2001b, 2001c, 2001d). The data simulated a 1- year (12 month) baseline period followed by a 5-year (60-month) remining period, with loadings measured once every month (also weekly, when the procedure required a period of accelerated monitoring). The evaluation examined the ability of a number of statistical procedures to react to various degrees of decrease and increase in loadings after baseline. The parametric procedures employed appropriate adjustments to the estimated variance to account for first-order serial correlation (assumed to be 0.5). The evaluation assumed that a minimum of 12 measurements of pollutant loads were made every year, once each month.

The ideal statistical procedure would always declare "not larger" when remining pollutant loadings are less than or equal to baseline loadings, and would always signal "larger" when remining loadings exceeded baseline. No such ideal procedure exists. Instead, the rate of signalling "larger" will increase as the average difference between baseline and remining loadings increases in magnitude. Statistical triggers may be "tuned", by choosing their numerical constants, so that a compromise is achieved between false alarms, that is, signalling "larger" when remining loadings are not larger than baseline loadings, and correct alarms, when remining loadings truly are greater.

The evaluations led to a choice of procedures and of numerical constants that achieve a reasonable balance between false alarms and correct alarms. This reasonable balance was considered to be achieved when a trigger produced the following results:

> (a) when there was no change in loadings from the baseline to the remining time period, the "false alarm rate" (type-I error rate) was not larger than that for the triggers used by the Commonwealth of Pennsylvania. Pennsylvania's trigger was used as a benchmark because of the demonstrated success of this approach (Hawkins 1994).

(b) when the mean pollutant load increased by one standard deviation after the baseline period, statistical power (probability of detecting the increase) was at least 0.75.

(c) when there was a decrease of 0.5 standard deviations in the mean loading after the baseline period, the"false alarm rate" was smaller than 5%.

(d) when the mean loading increased by 1 to 2 standard deviations after the baseline period, the "correct alarm rate" (power) was maximized (compared with other procedures).

Details of EPA's evaluation and comparisons of statistical procedures are provided in a separate document (U.S.E.P.A., 2001c). EPA reached the following conclusions about the statistical triggers based on these evaluations.

(1) The magnitude of serial correlation has a substantial effect on power. Statistical triggers that have reasonable power when there is no serial correlation could be unreasonable when there is substantial serial correlation, because they could then have very high rates of type I errors (false alarms). It was necessary to select numeric constants for the statistical triggers that are appropriate to data having autocorrelation. For evaluating and comparing statistical methods and triggers, EPA assumed a first-order autocorrelation coefficient of 0.5.

(2) To avoid false alarms, EPA determined that sequential exceedances of the Single Observation Trigger and accelerated monitoring were necessary. This method has long been used successfully in Pennsylvania's Remining Program. Specifically, the Single Observations Trigger requires the following: "If two successive monthly monitoring observations both exceed L, immediately begin weekly monitoring for four weeks (four weekly samples). If three or fewer of the weekly observations exceed L, resume monthly monitoring. If all four weekly observations exceed L, the baseline pollution loading has been exceeded."

(3) In Method 2, the Annual Comparison was set such that tables for the 99.9% level (alpha = 0.001) rather than the 95% level (alpha = 0.05) are to be used for the Wilcoxon-Mann-Whitney Test. When Type I Error rates of alpha = 0.05 or 0.01, the Wilcoxon-Mann-Whitney test in Method 2 had a high rate of declaring loadings to be larger than baseline when if fact, they were not larger.

(4) Method 1 and Method 2 were both designed as nonparametric rather than parametric procedures, with power comparable to that of a parametric procedure. Unlike a parametric method which would require log-transformation, the nonparametric methods accommodate zero flows (which may occur during remining) and negatively valued data (which may occur for net acidity) without requiring additional or complex modifications.

(5) EPA believes that the error rates and power of these triggers are acceptable in practice because best management practices (BMPs) reduced discharge loadings substantially. Hawkins (1994) reviewed the application of these triggers to remining operations in Pennsylvania, and concluded that the rates of triggering were low because remining usually reduced loadings substantially. EPA's BMP guidance manual includes an extensive analysis of remining discharges that supports this conclusion (EPA, 2001a). EPA concluded that the statistical triggers that Pennsylvania uses in its remining program are acceptable and effective and has used them as the basis for Method 1 with minor modifications to meet the criteria in (a) to (d). Method 1 herein follows the Pennsylvania triggers exactly except that a constant (1.815 = 1.96 * 1.25 / 1.35) is used in the formula for the Annual Procedure (see McGill, Tukey, and Larsen, 1978). Pennsylvania uses a more stringent number (1.58 = 1.7 * 1.25 / 1.35).

(6) The evaluation of the false alarm rate applies to a worst-case situation. The rate of declaring loadings to be larger than baseline when they are not is overstated by the evaluations (U.S.E.P.A., 2001c). It is evaluated in terms of the percentage of mines that would experience at least one determination that loadings exceed the baseline level over a period of five years (60 months), when in fact there has been no change from baseline. In practice, the area contributing to a discharge should be remined and regraded in less time, after which the discharge flow and loading will be substantially reduced. Thus the time period during which pollutant loadings are monitored for each discharge will usually

be shorter than five years. This in turn will mean lower percentages of false positives and false negatives than reported in Table 3.1a.

The power of statistical triggers for the final regulation is shown in Table 3.1a. The results show that Method 1 and Method 2 have comparable power to detect large increases (see columns '1σ' and '+2σ'). The main difference stems from the Monthly (Single Observation Limit) Test, which has higher false alarm percentages (see columns labeled '-0.5σ' and '0') when Method 1 is used.[1] Note that the Annual Comparison used without the Single Observation Limit Test would not have a high rate of detecting an increase of one standard deviation above baseline. Used in combination, the single observation and annual triggers provide power over 90% to detect substantial increases above baseline at least once during five years (Table 3.1a), although in practice the power may be smaller for reasons discussed above under (6).

[1]As explained in (5), EPA believes that the error rates and power of these triggers will be acceptable in practice because best management practices (BMPs) reduce discharge loadings substantially.

Table 3.1a. Statistical Triggers as Modified for Final Regulation: Percentage of Discharges Declared to Exceed Baseline Level (at least once during 5 years of simulated monthly monitoring) [1]

Annual Trigger	Single-Observation Trigger	Shift from Baseline to Remining Period [2]			
		-0.5 σ	0	1 σ	+2 σ
NA	Method 1	10 %	33 %	89 %	99 %
Method 1 (Multiplier = 1.96)	NA	3 %	11 %	59 %	94 %
Method 1 (Multiplier = 1.96)	Method 1	12 %	39 %	93 %	100 %
Method 1 (Multiplier = 1.96)	Method 2	7 %	29 %	91 %	100 %
NA	Method 2	5 %	22 %	86 %	100 %
Method 2 (α = 0.001)	NA	2 %	11 %	65 %	97 %
Method 2 (α = 0.001)	Method 2	7 %	28 %	91 %	100 %
Method 2 (α = 0.001)	Method 1	12 %	38 %	93 %	100 %

[1] Assumes monthly serial correlation of 0.5 for log(x), with x distributed lognormally. Percentages were rounded to the nearest 1%.

[2] The shift was scaled in terms of standard deviation units (σ = standard deviation)

3.2 Statistical Procedures for Calculating Limits from Baseline Data

The procedures to be used for establishing effluent limitations for pre-existing discharges at coal remining operations, in accordance with the requirements set forth in 40 CFR part 434, Subpart G; Coal Remining are presented below. The requirements specify that pollutant loadings of total iron, total manganese, total suspended solids, and net acidity in pre-existing discharges shall not exceed baseline pollutant loadings. The two alternative procedures described (Method 1 and Method 2) are applied to determine site-specific, baseline pollutant loadings, and to determine whether discharge loadings during coal remining operations have exceeded baseline loading. For each procedure, both a monthly (single-observation) test and an annual test are applied. In order to sufficiently characterize pollutant loadings during baseline determination and during each annual monitoring period, the regulations require that at least one sample result be obtained per month for a period of 12 months.

The calculations described are applied to pollutant loadings. Each loading value is calculated as the product of a flow measurement and pollutant concentration taken on the same date at the same discharge sampling point, using standard units of flow and concentration (to be determined by the permitting authority). For example, flow may be measured in cubic feet per second, concentration in milligrams per liter, and the pollutant loading calculated in pounds per year.

In the event that a pollutant concentration in the data used to determine baseline is lower than the daily maximum limitation established in Subpart C for active mine wastewater, the statistical procedures should not establish a baseline more stringent than the BPT and BAT effluent standards established in Subpart C. Therefore, if the total iron concentration in a baseline sample is below 7.0 mg/L, or the total manganese concentration is below 4.0 mg/L, the baseline sample concentration should be replaced with 7.0 mg/L and 4.0 mg/L, respectively, for the purposes of some of the statistical calculations. The substituted values should be used for all methods described in this section with the exception of the calculation of the interquartile range (R) in Method 1 for the annual trigger, and in Method 2 for the single observation trigger. The interquartile range (R) is the difference between the quartiles M_{-1} and M_1; these values should be

calculated using actual loadings (based on measured concentrations) when they are used to calculate R. This should be done in order to account for the full range of variability in the data.

3.2.1 Method 1

Method 1 is a modification of the methodology used by the Commonwealth of Pennsylvania. Computational details appear in Figure 3.2a. Pennsylvania's monthly and annual average checks can be described as follows:

Monthly (or single-observation maximum) check: A tolerance interval is estimated for the baseline loadings (for $n < 17$, the smallest and largest observations define the interval endpoints). The baseline upper bound (usually the maximum baseline loading) is the value of interest. Two consecutive exceedances of the upper bound trigger weekly monitoring. Four consecutive exceedances during weekly monitoring trigger a treatment requirement. Thus, six exceedances must occur consecutively before a treatment requirement is triggered.

Annual average check: A robust, asymptotic estimator[2] of a 95 percent confidence interval for the median is calculated for the baseline period and post-baseline periods; if the post-baseline interval exceeds the baseline interval, an exceedance is declared. This estimate is based upon McGill, Tukey, and Larsen (1978).

[2]Because loadings data for pre-existing discharges are highly asymmetric, and annual means and medians are likely to be somewhat asymmetrically distributed, EPA used an asymptotic approximation to develop the confidence intervals for the annual averages. However, the approximation results in confidence intervals that are symmetric rather than asymmetric. Thus, this approximation is expected to be accurate only for very large samples, because their means are approximately normally distributed (by the Central Limit Theorem). EPA has used this approximation for smaller samples because it provides reasonably good performance as demonstrated in simulations (see U.S.E.P.A., 2001c).

3.2.2 Method 2

Similarly to Method 1, Method 2 consists of two checks: an upper limit on single observations and an annual test of the mean or median. Computational details of Method 2 are provided in Figure 3.2b. The single-observation limit is a nonparametric estimate of the 99th percentile of loadings, developed using baseline data. The annual test of the average or median employs the nonparametric Wilcoxon-Mann-Whitney test.

3.2.3 Accelerated Monitoring

For Methods 1 and 2, triggered or accelerated monitoring is applied after two consecutive exceedances of the Single Observation Trigger L. If this occurs, weekly sampling must be commenced immediately. After four weekly samples are collected, the results should be compared to the Single Observation Trigger L. If three or fewer of the weekly observations exceed L, then monthly sampling can be resumed. However, if all four weekly observations exceed L, the baseline pollution loading has been exceeded.

Accelerated monitoring (if used as a condition or option for determining non-compliance) guards against a declaration of non-compliance on the basis of a transient exceedance, and provides a means to demonstrate continuing exceedances. It guards against the possibility of instituting expensive remedial measures when there was no continuing exceedance of baseline conditions.

Figure 3.2a: Method 1: The single-observation trigger is applied to each new measurement; the annual test is applied once a year, using all measurements for the past year

x_i = pollutant loading measurement (product of flow, concentration, and conversion factor)

n = number of x_i results in the baseline dataset

1. Single-observation trigger

Order all n baseline measurements such that $x_{(1)}$ is the lowest value, and $x_{(n)}$ is the highest.

If n<17, then:

The single-observation trigger will equal the maximum baseline value, $x_{(n)}$.

If n>16 then:

Calculate the sample median (M) of the baseline events:

If n is odd, then M equals $x_{(n/2+1/2)}$.

If n is even, then M equals $0.5*(x_{(n/2)}+ x_{(n/2+1)})$.

Calculate M_1 as the median between M and the maximum $x_{(n)}$.

Calculate M_2 as the median between M_1 and $x_{(n)}$.

Calculate M_3 as the median between M_2 and $x_{(n)}$.

Calculate M_4 as the median between M_3 and $x_{(n)}$.

The single-observation trigger L equals M_4.

If, during remining, two successive monthly observations exceed L, proceed immediately to weekly monitoring for four weeks (four weekly samples). If, during weekly monitoring, all four observations exceed L, declare exceedance of the baseline distribution.

2. Annual test

Calculate M and M_1 as described above.

Calculate M_{-1} as the median between the minimum $x_{(1)}$ and the sample median.

Calculate $R = (M_1 - M_{-1})$.

The subtle trigger (T) is calculated as:

$$T = M + \left[\frac{(1.815 * R)}{\sqrt{n}}\right]$$

Calculate M' and R' for a year's data during re-mining.

Calculate $T' = M' - (1.815*R')/(n')$.

If $T' > T$, conclude that the median loading during re-mining has exceeded the median loading during the baseline period, and declare an exceedance.

Figure 3.2b: Method 2: All three tests or limits are applied. The single-observation trigger is applied to each new measurement; the annual test is applied once a year, using all measurements for the past year

1. Single-observation trigger

Calculate M and M_1 as described in Method 1 (Figure 3.2a).
Calculate M_{-1} as the median between the minimum $x_{(1)}$ and the sample median.
Calculate R= (M_1 - M_{-1}).
Calculate the Single Observation Trigger as L = M_1 + (3 * R)

If, during remining, two successive monthly observations exceeds L, proceed immediately to weekly monitoring for four weeks (four weekly samples). If, during weekly monitoring, all four observations exceed L, declare exceedance of the baseline distribution.

2. Annual comparison [1]

Compare baseline year loadings with current annual loadings using the Wilcoxon-Mann-Whitney test [2] for two independent samples. Use a one-tailed test with alpha 0.001.

---

[1] Hirsch, R.M., and J.R. Stedinger. 1987. Plotting Positions for Historical Floods and Their Precision. Water Resources Research. Vol. 23, No.4:715-727.

[2] See Conover, W.J., 1980, Practical Nonparametric Statistics, 2nd ed., and other textbooks.

References

Brady, K.B.C., M.W. Smith, and J. Schueck (editors), 1998. *Coal Mine Drainage Prediction and Pollution Prevention in Pennsylvania*. Pennsylvania Department of Environmental Protection, publ. no. 5600-BK-DEP2256, October 1998.

Hawkins, J.W. 1994. *A statistical evaluation of remining abandoned coal mines to reduce the effluent contaminant load.* International Journal of Surface Mining, Reclamation and Environment 8: 101-109.

Hornberger, R.J. et al., 1990. *Acid Mine Drainage from Active and Abandoned Coal Mines in Pennsylvania.* Chapter 32 of Water Resources in Pennsylvania: Availability, Quality, and Management, Pennsylvania Academy of Science, pp. 432-451.

Loftis, J.C. and R.C.Ward, 1980. *Sampling Frequency Selection for Regulatory Water Quality Monitoring.* Water Resources Bulletin, Vol.16, No. 3, pp.501-507.

McGill, R., J.W. Tukey, and W.A. Larsen, 1978. *Variations of Box Plots*. The American Scientist, Vol. 32, No. 1, pp. 12-16.

U.S.E.P.A., 1993. *Statistical Support Document for Proposed Effluent Limitations Guidelines and Standards for the Pulp, Paper, and Paperboard Point Source Category*. EPA publ. no. 821/R-93-023.

U.S.E.P.A., 1999. Office of Water. *Coal Remining Database*: 61 State Data Packages, March 1999. (docket number DCN 3054)

U.S.E.P.A., 2001a. *Statistical Analysis of Abandoned Mine Drainage in the Assessment of Pollution Load*. EPA-821-B-01-014.

U.S.E.P.A., 2001b. "Serial Correlation of Coal Mine Discharge Loadings," memorandum in the rulemaking record (docket number DCN 3050).

U.S.E.P.A., 2001c. "Evaluation of Statistical Triggers," memorandum in the rulemaking record (docket number DCN 3051).

U.S.E.P.A., 2001d. "Distribution & Variability of Coal Mine Discharge Loadings," memorandum in the rulemaking record (docket number DCN 3049).